Methodological Issues & Concerns

H.T. McAdams

AccaMath Services
333 Ninth Street
Carrollton, IL 62016

Since I am unable to be here today for health reasons, I have prepared this summary of methodological issues that have concerned us and how we have dealt with them. As a result, I have reached certain convictions that I submit for your consideration.

Advantages of PCR+

- PCR+ offers maximum parsimony
- The best PCAR+ model may difer from any least-squares subset model based on property variables
- The fact that such a model contains all property variables makes explicit what is implicit in ordinary least-squares regression
- PCR+ provides estimates of regression coefficients for principal components, not property variables
- Both will be consistent if the number of eigenvectors and the number of selected property variables are the same
- "Pruning" of property variables from eigenvectors is a valid option but is not essential

Our first contention is that PCR+ models are the essence of brevity. This assertion is often countered by the objection that an eigenvector model retains all the original variables and that therefore no economy has been realized. We have shown, however, that in OLS models all variables are ALSO retained -- it is just that some of the variables are hidden.

In PCR+ we know explicitly how each transformed variable is defined. The model exhibits maximum parsimony because the dimensionality of the problem is appreciably reduced in the space of eigenvectors.

If it so happens that property variables can be "pruned" to conform to the number of eigenvectors in the model, the eigenvector coefficients can be simply transformed into the corresponding property-variable coefficients. This is not likely to happen, however, because dimension reduction is generally "steeper" for eigenvectors than for property variables.

Future Directions for Eigenfuels

- Selected issues on variable specifications – e.g., additized cetane
- Methodology for evaluating non-linear and interactive terms
- Select final emissions models for use in DOE refinery modeling on fuel reformulation
- Additional DOE/ORNL publications in late 2001 or 2002.

This chart summarizes the directions our work on eigenfuels is likely to take. There are some methodological issues we need to address, including how best to specify some variables like additized cetane and to evaluate a full range of non-linear and interactive terms. The issue on additized cetane is whether it is more effective to include total cetane in place of cetane difference and allow the eigenvector decomposition to sort out the difference between natural and total cetane.

Having done this, our next major milestone is to select a final emissions model for use in DOE refinery analyses related to diesel fuel reformulation.

We expect to release additional publications on this work in late 2001 or early 2002.

Predictive Ability of Models

- R-Square measures only how well a regression equation fits the data set: it does not necessarily insure good predictions elsewhere in property space.
- Validation testing using a split sample may not guarantee good prediction either: randomly selected samples are likely to exhibit the same aliasing structure, so that agreement of the samples is somewhat pre-ordained.
- Property-based regression produces many models that are nearly equal in performance in the discrete space of observations.
- PCR+ models obey a hierarchy that makes for unique choices.

We come, now, to question the venerable concept of R-Square. The fact is that R-Square, the so-called Coefficient of Determination, is based ENTIRELY on the observed data in the dataset. Therefore, it is a measure of degree of fit in a DISCRETE environment consisting of a disjoint and unconnected set of points in predictor space. Theoretically, there exists an infinite number of CONTINUOUS functions that can fit the data points equally well. Which one is "correct" is not a foregone conclusion.

Many subset models based on fuel properties exhibit nearly equal R-Square values and even yield nearly the same point-by-point predictions so long as these predictions are limited to the dataset. Their predictions in continuous space can differ materially, however, a point that requires no proof when it is realized that the subset models contain DIFFERENT variables and that a retained variable can not explain what a missing variable could.

Subset models in property space exhibit no orderly basis for choice, whereas eigenvector models are strictly ordered so that optimum choice is evident.

Other Aspects of Modeling

- Regression models are artifacts of the statistical tests employed to include or exclude terms.
- The 0.05 significance level is not sacrosanct; considerations of the power of a test of significance needs to be taken into account.
- Measures of the magnitudes of effects should supplement measures of statistical significance; we call these measures of substantiality.

There are other aspects of modeling that are of concern. These have been detailed in our previous publications and are listed here for completeness.

In summary, we believe that our approach constitutes a different PARADIGM for model building and that it is well supported by both theoretical and practical considerations.

Eigenvector Modeling using PCR+

- Every dataset has its own set of independent vectors that eliminate such aliasing.
- These eigenvectors are the "true" variables and are the only ones capable of providing an unambiguous model.
- PCR+ is an extension of the method described in the literature as PCR. Its distinction consists in the way that vectors are selected for inclusion in the model.

It is at this point that the notion of eigenvector modeling comes into play. The fact that every dataset has an orthogonal basis is at the root of the methodology that we have referred to as PCR+. It uses the methodology of Principal Components Regression (PCR) except for the way in which one selects the set of eigenvectors to retain in a subset model.

For some reason that I cannot understand, data analysts have used the eigenvalues of the design matrix (X-matrix) as the basis for including or excluding eigenvectors. As late as 1998 critics (see Hadi and Ling) were still pointing out that such an approach could lead to miserable failure.

In reality, the columns of numbers in a design matrix consisting of the principal components is just that: columns of numbers. If they were labeled X1, X2, X3, ... Xn and given as a textbook "exercise for the student," no one would ever suspect their shady past. Given a response vector Y, the student would faithfully do what is expected: derive a multiple regression equation, perform some tests of significance and publish the results on the Internet.

In short, PCR+ employs the methodology of least squares regression and benefits from all of its characteristics, such as unbiasedness and minimum variance estimation. Moreover, it is not subject to bias, as is stepwise regression, when terms are removed from the model because the alias matrix is null. Therefore, coefficients are invariant when terms are added to or removed from the model.

The eigenVALUES pertain only to the variation AMONG THE PREDICTOR VARIABLES and have nothing to do with the response.

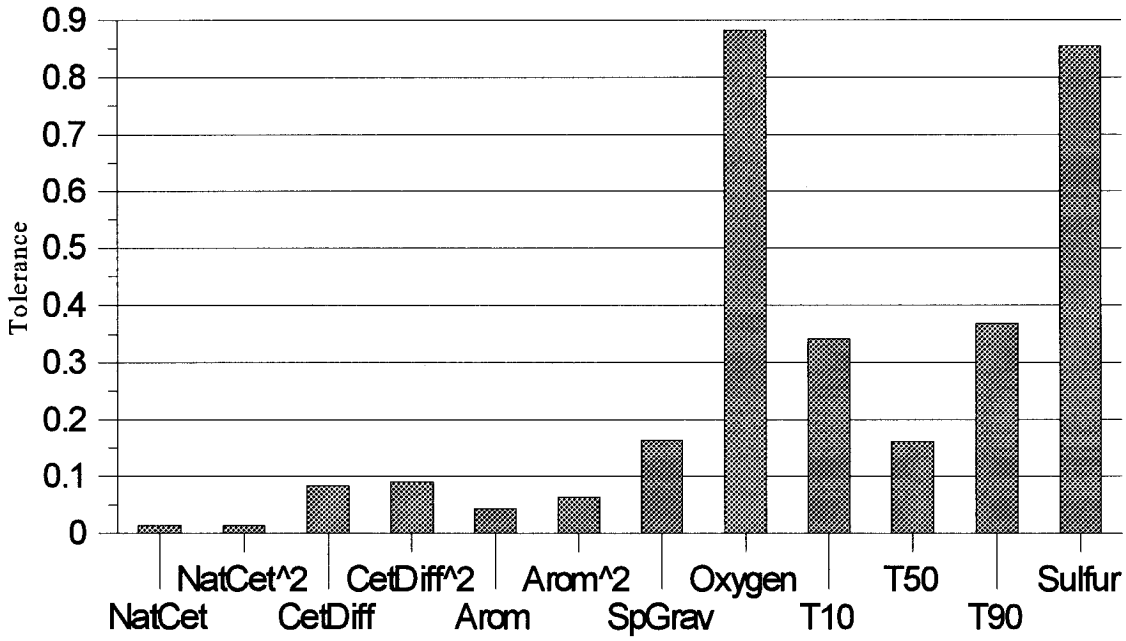
Aliasing Among Variables

- Except for balanced experimental designs, every dataset has its own peculiar interdependencies.
- These interdependencies produce aliasing among predictor variables.
- Selection of variables from such a dataset only changes the aliasing; it does not eliminate it.
- Variables, therefore, have chameleon characteristics and can not be interpreted as representing what their names imply.

My first concern is with what we have come to call "aliasing" among interdependent variables. It is, I believe, the primary source of many if not most of the difficulties analysts face in formulating a regression equation to represent a given dataset. It is quite important, therefore, that this concept and its implications be fully understood.

I believe that conventional approaches to modeling interdependent data represent, for the most part, means to CIRCUMVENT rather than REMEDY this defect. The next few slides will address this issue.

Interdependence of Fuel Variables



My demonstration is based on engine-corrected NOx data for technology group T, for which the seriousness of interdependence is illustrated by this slide.

Because of the interdependencies among the predictor variables, the regression coefficient for a particular predictor -- say, aromatics -- represents regression on ONLY that part of the predictor that is NOT explained by other predictor variables. That part is measured by a quantity called TOLERANCE. It is one minus the R-Square obtained when that variable is regressed on all the other predictors.

In this chart it is evident that several of the predictors have very low tolerances and that only two, oxygen and sulfur content, are relatively independent of the others.

Effect of Variable Selection on Regression Coefficients

Fuel Property	Before Selection Coeff	t	After Selection Coeff	t	
Nat Cetane	-0.0077	0.58			
Nat Cetane^2	-0.0042	0.31			
Cet Diff	-0.0289	6.15*	-0.0273	5.76*	
Cet Diff^2	-0.0127	2.78*	0.0122	2.62*	
Aromatics	0.0324	5.24*	0.0248	10.6*	<--
Aromatics^2	-0.0098	1.81			
Spec Gravity	0.0106	3.20*	0.0203	8.61*	<--
Oxygen Content	0.0053	3.68*	0.0055	3.74*	
T10	0.0104	4.52*	0.0103	4.82*	
T50	-0.0104	3.05*	-0.0173	7.57*	<--
T90	0.0021	0.95			
Sulfur	-0.0021	1.43			

We begin our demonstration by performing an ordinary regression in which all 12 fuel-property variables are included as predictors. Because of the fact that the emissions are deviations from a mean value, they sum to zero and have zero intercept.

As is usually done, those variables satisfying the 0.05 significance level are retained; all others are rejected. In this case 7 predictors are retained. The result is one of the 4095 possible subset models.

I call your attention to three variables -- aromatics, specific gravity, and T50 -- that appear to be much more significant in the subset model than in the full model.

What Caused the Change in Aromatics, Specific Gravity and T50?

	REGRESSION COEFFICIENTS			
Fuel Property	All Variables	Variables Eliminated	Variables Retained	
Nat Cetane	-0.0077	-0.0077		
Nat Cetane^2	-0.0042	-0.0042		
Cet Diff	-0.0289		-0.0273	
Cet Diff^2	-0.0127		0.0122	
Aromatics	0.0324		0.0248	<--
Aromatics^2	-0.0098	-0.0098		
Spec Gravity	0.0106		0.0203	<--
Oxygen Content	0.0053		0.0055	
T10	0.0104		0.0103	
T50	-0.0104		-0.0173	<--
T90	0.0021	0.0021		
Sulfur	-0.0021	-0.0021		

Just what caused the change in regression coefficients and their apparent significance for these three variables? Evidently these variables benefited from the deletion of the 5 variables: Nat Cetane, Nat Cetane^2, Aromatics^2, T90 and Sulfur. Exactly HOW this benefit comes about will be shown in the next two slides.

The change in the value of the retained coefficients can be computed explicitly by means of a quantity called the ALIAS matrix or, for reasons to be shown later, sometimes also called the BIAS matrix. I will not attempt here and now to show the theory underlying this computation. Its effect, however, is that each of the rejected coefficients is given a weight specific to each of the retained coefficients. The weighted sum of the rejected coefficients is then added to the retained coefficient as initially computed when all 12 variables were in the model. It will be seen that the added portion is exactly equal to the difference between the coefficients for the subset model and for the full model.

Aliasing of Aromatics to Other Variables

Aromatics Coefficent (Full Model):	0.0324
+ contributions from	
Natural Cetane	0.0016
Natural Cetane^2	0.0007
Aromatics^2	-0.0103
T90	0.0005
Sulfur content	-0.0001
=	0.0248
Aromatics Coefficient (Subset Model)	0.0248

This chart shows how the aromatics coefficient changes from 0.0324 in the full model to 0.0248 in the subset model. Note that the major source of the change is a contribution from the Aromatics^2 term. Thus, what we originally thought was a linear aromatics effect is now "aliased" with its square term.

Aliasing of Specific Gravity

Spec Grav Coefficent (Full Model):	0.0106
+ contributions from	
Natural Cetane	0.0058
Natural Cetane^2	0.0034
Aromatics^2	0.0013
T90	-0.0004
Sulfur content	-0.0004
=	0.0203
Spec Grav Coefficient (Subset Model)	0.0203

The case is not so simple for specific gravity.

Here, the coefficient for specific gravity is almost doubled in going from the full model to the subset model. This change is primarily atrributable to natural cetane and its square term. What we originally thought was the effect of specific gravity, therefore, is now aliased with the effect of natural cetane and its square.

The generalization to be made here is that similar modifications in coefficients will be made for ANY subset of the predictor variables. Relative to the full model, the coefficients in the subset model may be said to be BIASED, and it is for this reason that the transforming matrix is sometimes called the BIAS matrix.

It should be noted, also, that the coefficients for ANY SUBSET MODEL can be computed directly from the coefficients for the full model, without performing the least-square procedure for the subset.